

Conversations with AutoTutor Help Students Learn

Arthur C. Graesser¹

Published online: 19 January 2016

© International Artificial Intelligence in Education Society 2016

Abstract AutoTutor helps students learn by holding a conversation in natural language. AutoTutor is adaptive to the learners' actions, verbal contributions, and in some systems their emotions. Many of AutoTutor's conversation patterns simulate human tutoring, but other patterns implement ideal pedagogies that open the door to computer tutors eclipsing human tutors in learning gains. Indeed, current versions of AutoTutor yield learning gains on par with novice and expert human tutors. This article selectively highlights the status of AutoTutor's dialogue moves, learning gains, implementation challenges, differences between human and ideal tutors, and some of the systems that evolved from AutoTutor. Current and future AutoTutor projects are investigating three-party conversations, called *trialogues*, where two agents (such as a tutor and student) interact with the human learner.

Keywords AutoTutor · Conversational agents · Trialogues · Intelligent tutoring systems

The Tutorial Dialogue of AutoTutor

This article reflects on a paper published in 2001 on “Teaching Tactics and Dialogue in AutoTutor”, coauthored by Graesser et al. (2001). *AutoTutor* is a pedagogical agent that holds a conversation with students in natural language and simulates the dialogue moves of human tutors as well as ideal pedagogical strategies (Graesser et al. 2004; Graesser et al. 2008; see Nye et al. 2014, for an in depth history of 17 years of AutoTutor). My colleagues and I were inspired by the notion that there is something about conversation mechanisms in a tutoring session that help people learn (Graesser et al. 1995). And indeed, untrained tutors do help students learn better than classroom

Submission to Special Issue of *International Journal of Artificial Intelligence in Education* on 25th Anniversary.

✉ Arthur C. Graesser
graesser@memphis.edu

¹ Department of Psychology & Institute for Intelligent Systems, University of Memphis, 202 Psychology Building, Memphis, TN 38152-3230, USA

interactions and various other ecological controls (Graesser et al. 2011). We were also inspired by the notion that some of the discourse moves of tutors could be improved if they were guided by ideal pedagogical principles. So a combination of natural discourse interaction and ideal tutor moves would be the magic formula to improve student learning.

We wrestled with the possibility that ideal computer tutoring moves may be different sometimes than normal conversational moves of human tutors. For example, our analyses of human tutors revealed that they are prone to follow principles of conversational politeness so they are reluctant to give negative feedback when a student's contribution is incorrect or vague (Graesser et al. 1995; Person et al. 1995). Accurate feedback sometimes needs to be sacrificed in order to promote confidence and self-efficacy in the student (Lepper and Woolverton 2002). However, many students expect the computers to be accurate rather than polite. Consequently, there is a trade-off between feedback accuracy and the promotion of politeness or self-esteem. There also appeared to be illuminating differences in the pragmatic ground rules of communication with computers versus humans (Person et al. 1995). Given the various trade-offs and incompatible predictions, we envisioned a program of research to investigate the impact of specific tutoring strategies and conversation patterns on student learning and motivation. This program of research continues to evolve among colleagues investigating automated tutorial dialogue both in Memphis (Nye et al. 2014; Rus et al. 2013) and other labs (e.g., Dzikovska et al. 2014; Johnson and Lester 2015; Ward et al. 2013).

The AutoTutor project was launched in 1997 at a point in history when animated conversational agents emerged and penetrated learning environments. The agents were computerized talking heads or embodied animated avatars that generate speech, actions, facial expressions, and gestures. Some of the agents were very rigid and scripted, whereas AutoTutor attempted to adapt to the knowledge states, verbosity, and the emotional states of the learner. AutoTutor was indeed successful in tracking the student's knowledge states and adaptively generating dialogue moves (Graesser et al. 2004; Jackson and Graesser 2006; Nye et al. 2014; VanLehn et al. 2007). We also developed an affect-sensitive AutoTutor that responded intelligently to the emotions of the student, such as confusion, frustration, and boredom (D'Mello and Graesser 2012). The power of conversational agents is that they can precisely specify what the agent expresses and does under specific conditions, whereas humans could never exhibit such precision. Agents can guide the learner on what to do next, deliver didactic instruction, hold collaborative conversations, and model ideal behavior, strategies, reflections, and social interactions. Pedagogical agents have become increasingly popular in contemporary adaptive learning environments (*DeepTutor*, Rus et al. 2013; *Betty's Brain*, Biswas et al. 2010; *iSTART*, McNamara et al. 2006; *Crystal Island*, Rowe et al. 2010; *Guru Tutor*, Olney et al. 2012; *Operation ARIES*, Millis et al. 2011), just to name a few systems. These systems have covered topics in STEM (physics, biology, computer literacy), reading comprehension, scientific reasoning, and other domains and skills.

AutoTutor and these other systems with pedagogical agents have helped students learn compared to various control conditions. In the case of AutoTutor, reports covering multiple studies have reported average learning gains that vary between 0.3 sigma (Nye et al. 2014) and 0.8 (Graesser et al. 2008) when compared to reading text for an equivalent amount of time; the effect sizes are substantially higher in comparisons with

pre-tests and no-study controls (Graesser et al. 2004; VanLehn et al. 2007). Human tutors have not differed greatly from AutoTutor and other ITS's with natural language interaction in experiments that provide direct comparisons with trained human tutors (Olney et al. 2012; VanLehn 2011; VanLehn et al. 2007). For example, in a direct comparison between AutoTutor and 1-to-1 human tutoring with experienced tutors in computer-mediated conversations (either typed or spoken), the learning gains were virtually equivalent on the topic of Newtonian physics (VanLehn et al. 2007). Given these encouraging results from human and computer tutoring, we investigated what it is about conversation that helps student learning and motivation.

Conversation Patterns in AutoTutor and Human Tutors

We conducted a series of experiments that attempted to identify the features of AutoTutor that might account for improvements in learning (Graesser et al. 2004, 2008; Kopp et al. 2012; VanLehn et al. 2007). It is beyond the scope of this article to cover all of these features, but a few are particularly noteworthy. One noteworthy finding is that it is not the talking head that accounts for most of the improvement, but rather the content of what the agent says and the student says. The talking head has only a small advantage over the agent's conveying its dialogue moves in print or spoken modalities. Learning from AutoTutor is not appreciably different from conditions where the learner is guided to read small snippets of text or summaries of a solution at opportunistic points in time. From the standpoint of student input modality, learning is no different when students express their contributions in speech or keyboard (D'Mello et al. 2011). Simply put, it is the content that matters: What gets expressed at the right time in a conversation?

Another noteworthy conclusion is that we were impressed with the robustness of the core conversation mechanisms in both AutoTutor and most human tutoring. As mentioned earlier, many of the core conversation mechanisms in AutoTutor are similar to human tutoring. We documented major conversation mechanisms of human tutors who tutored middle school children in mathematics and college students in research methods. The detailed anatomy of human tutoring was based on near 100 tutoring sessions that were videotaped, transcribed, and analyzed in depth (Graesser et al. 1997; Graesser and Person 1994; Graesser et al. 1995; Person et al. 1994; Person et al. 1995). In particular, one discourse mechanism in both AutoTutor and human tutoring is called *expectation & misconception-tailored dialogue* (EMT dialogue). The human tutors anticipate particular correct answers (called expectations) and particular misunderstandings (misconceptions) when they ask the students challenging questions (or problems) and track the students' answers. As the students express their answers, which are distributed over multiple conversational turns, their contributions are compared with expectations and misconceptions through semantic pattern matching. The tutors give feedback to the students' answers with respect to matching the expectations or misconceptions. Some feedback is short, consisting of positive, neutral, or negative expressions either in words, intonation, or facial expressions. After the short feedback, the tutor tries to lead the student to express the expectations (good answers) through multiple dialogue moves, such as *pumps* ("What else?"), *hints*, or *prompts* to get the students to express specific words. When the student fails to answer the question correctly, the tutor contributes information as *assertions*. The pump-hint-prompt-assertion cycles are implemented in AutoTutor (and are frequent in

human tutoring, Graesser et al. 1995) to extract or cover particular sentence-like expectations. Eventually, all of the expectations are covered and the exchange is finished for the main question or problem.

It is feasible to implement EMT dialogue computationally because it relies on semantic pattern matching and attempts to achieve pattern completion (through hints and prompts). This is a simpler mechanism than interpreting natural language from scratch, which is beyond the boundaries of reliable natural language processing. EMT dialogue is not only frequent in human tutoring but creates reasonably smooth conversations in AutoTutor and helps students learn. Interestingly, human tutors rarely use sophisticated tutoring strategies that are difficult to implement on computer, such as bona fide Socratic tutoring, modeling-scaffolding-fading, building on prerequisites, and dialogue moves that scaffold metacognitive strategies (Cade et al. 2008; Graesser et al. 1995). Automated computer tutors will possibly show major advantages over human tutors when the systems can reliably implement these more sophisticated strategies.

AutoTutor successfully implemented nearly all of the conversational mechanisms of human tutors, but one notable exception is that it could not handle most of the student questions. Student questions are infrequent in most classroom and tutoring environments because the teacher or tutor tends to control the agenda (Graesser et al. 1995). However, when students do ask questions, the relevance and correctness of the answers is disappointing in AutoTutor, as it is in other automated environments. We have had to implement diversionary tactics to handle the students questions, such as “How would you answer that question?” or “AutoTutor cannot answer that question now.” As a consequence, the frequency of student questions unfortunately extinguishes quickly in tutoring sessions with AutoTutor (Graesser and McNamara 2010).

We continued to question the use of human tutors, even expert tutors, as the gold standard in the design of AutoTutor. We identified a number of blind spots and questionable tactics of human tutors (Graesser et al. 2011) that could potentially be improved by incorporating ideal tutoring strategies. For example, tutors are prone to give a summary recap of a solution to a problem, or answer to a difficult question, that required many conversational turns. It would be better to sometimes have the student give the summary recap in order to promote active student learning, to encourage the student to practice articulating the information, or to allow the tutor to diagnose remaining deficits. As another example, tutors often assume that the student understands what the tutor expresses in an exchange whereas students often do not understand, even partially. Indeed, there often is a large gulf between the knowledge of the student and that of the tutor. It sometimes would be better for the tutor to ask follow up questions to verify the extent to which the student understands what the tutor is attempting to communicate. Ideal tutoring strategies are needed to augment or replace some of the typical conversation patterns in human tutoring.

One of the pervasive challenges throughout the development of AutoTutor and subsequent learning environments has been optimizing the semantic match scores between the students’ verbal contributions and AutoTutor’s anticipated answers (both the expectations and misconceptions). The student’s contributions over dozens of conversational turns in a single dialogue are constantly compared semantically with the set of expectations and misconceptions. There is a speech act classifier that segments the student’s verbal input within a turn into speech acts and assigns each speech act to a category, such *question*, *statement*, *meta-cognitive expression* (e.g., “I do not know”) or *short response*, as designated in the Dialogue Advancer Network

(Graesser et al. 2001). The statements are the only speech acts that are compared with the expectations and misconceptions through semantic matching algorithms. An expectation (or misconception) is considered covered if it meets or exceeds some threshold parameter for matching.

We have evaluated many semantic matchers over the years. The best results are a combination of latent semantic analysis (LSA) (Landauer et al. 2007), frequency weighted word overlap (rarer words and negations have higher weight), and regular expressions. In fact, LSA plus regular expressions have had high reliability scores in comparisons with human experts versus pairs of human experts (Cai et al. 2011). Interestingly, syntactic computations did not prove useful in these analyses because a high percentage of the students' contributions are telegraphic, elliptical, and ungrammatical. Researchers who have developed tutorial dialogue systems with deep syntactic parsers (e.g., BEETLE II, Dzikovska et al. 2014) routinely point out the limitations of syntactic parsers when there are low quality language contributions of students.

We learned, after many years, that a semantic match algorithm with impressive fidelity will not necessarily go the distance in meeting the students' wishes. There are two problems that continue to haunt us. The first problem addresses the students' standards on what it means to cover a sentence-like answer correctly. If a good answer has four content words (A, B, C, D) that ideally are expressed, the students want full credit if they can express only one or two of the distinctive words (e.g., A and B). They get frustrated when their partial answers only get neutral or negative feedback from the tutor; students think they have covered the sentence answer but AutoTutor does not score it as covered unless the students express the remaining words (C and D). The students assume that the assumed shared knowledge should be sufficient to fill in the remaining words, but AutoTutor wants to see a more complete answer articulated. The second problem addresses the semantic blur that invariably occurs between expectations and misconceptions when the algorithms rely on statistical algorithms like LSA, word overlap and regular expressions. Students may get negative feedback when their statements match a misconception more than an expectation; or positive feedback when they express something erroneous. This semantic blur produces inaccurate feedback which can end up confusing or frustrating the student. Although we do everything we can to engineer the content and threshold parameters, these errors still occasionally occur because of the vagueness of language. One practical solution to this challenge is to have AutoTutor give neutral short feedback after these uncertain or borderline semantic matches so that the student is not misled or frustrated when the semantic matches are imperfect. Another approach is to provide more discriminating hints and prompts when there is a semantic blur between expectations and misconceptions. The hints and prompts would more cleanly differentiate a correct expectation versus a misconception.

Future AutoTutor Directions and Trialogues

Many spinoffs from AutoTutor have been developed after its inception in 1997 and the publication of Graesser et al. (2001). Nye et al. (2014) reported that dozens of systems

have evolved from AutoTutor in the Institute for Intelligent Systems at the University of Memphis. These systems have covered many STEM topics, reading comprehension, writing, and scientific reasoning, with names like DeepTutor, GuruTutor, GnuTutor, AutoMentor, iDRIVE, iSTART, Writing-Pal, Operation ARIES (and ARA). A recent system has integrated AutoTutor with ALEKS in mathematics, a system commercialized by McGraw-Hill that has helped middle school students in the Memphis area (Hu et al. 2012). The Memphis team has recently started developing AutoTutor for basic electronics and electricity in an ElectronxTutor that is funded by Office of Naval Research. The suite of AutoTutor applications is starting to cover a large curriculum landscape. Researchers at other universities, businesses, and organizations are increasingly licensing the AutoTutor Script Authoring Tool (Cai et al. 2015) to develop their own content and integrate it with our generic AutoTutor Conversation Engine (ACE). For example, Wolfe et al. (2015) used the AutoTutor authoring tools to develop a website on genetic risk factors for breast cancer, called BRCA, and reported learning gains above the existing web site on the same topic. Educational Testing Service is licensing ASAT for assessment on a variety of competencies (English Language Learning, science, mathematics) in the context of virtual worlds with agents (Zapata-Rivera et al. 2015). The Army Research Laboratory has incorporated AutoTutor in its open source Generalized Intelligent Framework for Tutoring (GIFT, Sottolare et al. 2013). AutoTutor is growing further as it migrates to new systems with new names and applications.

In recent years we have developed dialogues, which involve the human interacting with two agents, typically a student agent and a tutor agent in 3-party conversations (Graesser et al. 2014; Graesser et al. 2015a, b; Millis et al. 2011). Two agents add considerable benefits theoretically because the two agents can model successful conversational interactions, such as asking good questions and receiving good answers (Gholson et al., 2009) or staging arguments that create cognitive disequilibrium, productive confusion, and deeper learning (D’Mello et al. 2014; Lehman et al. 2013). The dialogues can help rectify some of the problems previously discussed on AutoTutor dialogues. For example, when the human’s answer is incomplete, the student agent can fill in the missing words and articulate a more complete answer; this not only models good answers but also circumvents any negative short feedback to the human.

Graesser et al. (2015b) identified seven dialogue designs that can be used in learning environments. The two agents in each design can take on different roles, but typically one is a tutor and the other a student peer.

- (1) *Vicarious learning with human observer.* Two agents interact and model ideal behavior, answers to questions, or reasoning.
- (2) *Vicarious learning with limited human participation.* The same as #1 except that the agents occasionally turn to the human and ask a prompt question, with a yes/no or single-word answer.
- (3) *Tutor agent interacting with human and student agent.* There is a tutorial dialogue with the human, but the student agent periodically contributes and receives feedback.

- (4) *Expert agent staging a competition between the human and a peer agent.* There is a competitive game between the human and peer agent, with the expert agent organizing the event.
- (5) *Human teaches/helps a student agent with facilitation from the tutor agent.* As the human tries to help the peer agent, the tutor agent rescues a problematic situation.
- (6) *Human interacts with two peer agents that vary in proficiency.* The peer agents can vary in knowledge and skills.
- (7) *Human interacts with two agents expressing contradictions, arguments, or different views.* The discrepancies between agents stimulate cognitive disagreement, confusion, and potentially deeper learning.

Our current hypothesis is that these seven trialogue designs should be adaptively administered, depending on the student's knowledge and other psychological attributes. The vicarious learning designs (1 and 2) are appropriate for learners with limited knowledge, skills, and actions, whereas designs 5 and 7 are suited to the more capable students attempting to achieve deeper knowledge. Design 4 is motivating for learners by virtue of the game competition. Research needs to be conducted to assess empirically the conditions under which different trialogue designs facilitate learning and motivation.

Trialogues have been routinely incorporated in our recent AutoTutor applications. Scientific reasoning is the focus in an instructional game called *Operation ARIES!* (Millis et al. 2011), which was subsequently commercialized by Pearson Education as *Operation ARA* (Halpern et al. 2012). ARIES is an acronym for **A**cquiring **R**esearch **I**nvestigative and **E**valuative **S**kills whereas ARA is an acronym for **A**cquiring **R**esearch **A**cumen. Agent dialogues are currently being developed in computer interventions to train comprehension strategies for adults with reading difficulties in the Center for the Study of Adult Literacy (CSAL, <http://csal.gsu.edu/content/homepage>). Interestingly, some trialogue designs have always been used in McNamara's iSTART trainer for reading comprehension (McNamara et al. 2006). ETS is currently using dialogues for assessment and is licensing our ASAT and ACE facilities for that purpose (Zapata-Rivera et al. 2015).

It is of course possible to build systems with more than two agents and more than one human. One can imagine communities of humans and cyber agents interacting in varying numbers. The cyber agents will need conversation mechanisms that are adaptive and flexible in a similar vein as AutoTutor dialogues and dialogues. At that point we enter the arenas of collaborative problem solving (Fiore et al. 2010; Graesser et al. 2015a) and computer supported collaborative learning (Dillenbourg 1999; Rosé et al., 2008). These are two areas on our horizon during the next decade.

Acknowledgments This research was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918, ALT-0834847, DRK-12-0918409, 1108845), the Institute of Education Sciences (R305H050169, R305B070349, R305A080589, R305A080594, R305G020018, R305C120001), Army Research Lab (W911NF-12-2-0030), and the Office of Naval Research (N00014-00-1-0600, N00014-12-C-0643). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, physics, and education at University of Memphis (visit <http://www.autotutor.org>, <http://emotion.autotutor.org>, <http://fedex.memphis.edu/iis/>). Requests for reprints should be sent to Art Graesser, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230, graesser@memphis.edu.

References

- Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology-Enhanced Learning*, 5, 123–152.
- Cade, W., Copeland, J., Person, N., & D’Mello, S. K. (2008). Dialogue modes in expert tutoring. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the ninth international conference on intelligent tutoring systems* (pp. 470–479). Berlin, Heidelberg: Springer-Verlag.
- Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., et al. (2011). Trialog in ARIES: user input assessment in an intelligent tutoring system. In W. Chen, & S. Li (Eds.), *Proceedings of the 3rd IEEE international conference on intelligent computing and intelligent systems* (pp. 429–433). Guangzhou: IEEE Press.
- Cai, Z., Graesser, A. C., & Hu, X. (2015). ASAT: AutoTutor script authoring tool. In R. Sottolare, A. C. Graesser, X. Hu, & K. W. Brawner (Eds.), *Design recommendations for intelligent tutoring systems: authoring tools* (vol. 3,). Orlando, FL: Army Research Laboratory.
- D’Mello, S. K., & Graesser, A. C. (2012). AutoTutor and affective AutoTutor: learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2, 1–39.
- D’Mello, S., Dowell, N., & Graesser, A. C. (2011). Does it really matter whether students’ contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied*, 17, 1–17.
- D’Mello, S., Lehman, B., Pekrun, R., & Graesser, A. C. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153–170.
- Dillenbourg, P. (1999). *Collaborative learning: cognitive and computational approaches. Advances in learning and instruction series*. New York, NY: Elsevier Science INC..
- Dzikovska, M., Steinhäuser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24, 284–332.
- Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: predicting processes in complex collaborative contexts. *Human Factors*, 52, 203–224.
- Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., et al. (2009). Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and Newtonian physics. *Instructional Science*, 37, 487–493.
- Graesser, A. C., & McNamara, D. S. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist*, 45, 234–244.
- Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31, 104–137.
- Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9, 495–522.
- Graesser, A. C., Bowers, C. A., Hacker, D. J., & Person, N. K. (1997). An anatomy of naturalistic tutoring. In K. Hogan, & M. Pressley (Eds.), *Scaffolding student learning: instructional approaches and issues* (pp. 145–184). Cambridge, MA: Brookline Books.
- Graesser, A. C., Person, N., Harter, D., & the Tutoring Research Group (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257–279.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: a tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36, 180–193.
- Graesser, A. C., Jeon, M., & Duffy, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes*, 45, 298–322.
- Graesser, A. C., D’Mello, S. K., & Cade, W. (2011). Instruction based on tutoring. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 408–426). New York: Routledge Press.
- Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, 23, 274–280.
- Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M. (2015a). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills*. Berlin: Springer Publishers (in press).

- Graesser, A.C., Forsyth, C., & Lehman, B. (2015b). Two heads may be better than one: learning from computer agents in conversational dialogues. *Teachers College Record* (in press).
- Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: a computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity*, 7, 93–100.
- Hu, X., Craig, S. D., Bargagliotti, A. E., Graesser, A. C., Okwumabua, T., Anderson, C., et al. (2012). The effects of a traditional and technology-based after-school program on 6th grade students' mathematics skills. *Journal of Computers in Mathematics and Science Teaching*, 31, 17–38.
- Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: an intelligent tutoring system. *Revista Signos*, 39, 31–48.
- Johnson, W.L. & Lester, J.C. (2015). Twenty years of face-to-face interaction with pedagogical agents. *International Journal of Artificial Intelligence in Education* (in press).
- Kopp, K., Britt, A., Millis, K., & Graesser, A. (2012). Improving the efficiency of dialogue in tutoring. *Learning and Instruction*, 22, 320–330.
- Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Lehman, B., D'Mello, S. K., Strain, A., Mills, C., Gross, M., Dobbins, A., et al. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education*, 22, 85–105.
- Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: impact of psychological factors on education* (pp. 135–158). Orlando, FL: Academic Press.
- McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, 34, 147–171.
- Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp. 169–196). London, UK: Springer-Verlag.
- Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: a review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24, 427–469.
- Olney, A., D'Mello, S. K., Person, N., Cade, W., Hays, P., Williams, C., et al. (2012). Guru: a computer tutor that models expert human tutors. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of intelligent tutoring systems (ITS) 2012* (pp. 256–261). Berlin, Germany: Springer.
- Person, N. K., Graesser, A. C., Magliano, J. P., & Kreuz, R. J. (1994). Inferring what the student knows in one-to-one tutoring: the role of student questions and answers. *Learning and Individual Differences*, 6, 205–229.
- Person, N. K., Kreuz, R. J., Zwaan, R., & Graesser, A. C. (1995). Pragmatics and pedagogy: conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction*, 13, 161–188.
- Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3, 237–271.
- Rowe, J., Shores, L. R., Mott, B., & Lester, J. (2010). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 19, 166–177.
- Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent systems with conversational dialogue. *AI Magazine*, 34, 42–54.
- Sottolare, R., Graesser, A., Hu, X., & Holden, H. (Eds.) (2013). *Design recommendations for intelligent tutoring systems: learner modeling* (vol. 1,). Orlando, FL: Army Research Laboratory.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist*, 46, 197–221.
- VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31, 3–62.
- Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013). My science tutor: a conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105, 1115–1125.
- Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., & Weil, A. M. (2015). Efficacy of a web-based intelligent tutoring system for communicating genetic risk of breast cancer: a fuzzy-trace theory approach. *Medical Decision Making*, 35, 46–59.
- Zapata-Rivera, D., Jackson, T., & Katz, I.R. (2015). Authoring conversation-based assessment scenarios. In R. Sottolare, A.C. Graesser, X. Hu, & K.W. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools* (Vol.3). Orlando, FL: Army Research Laboratory.